

# Weekly Report

Period: 2017/8/28-2016/9/3

Reporter: Li Zongzhuang

## Done

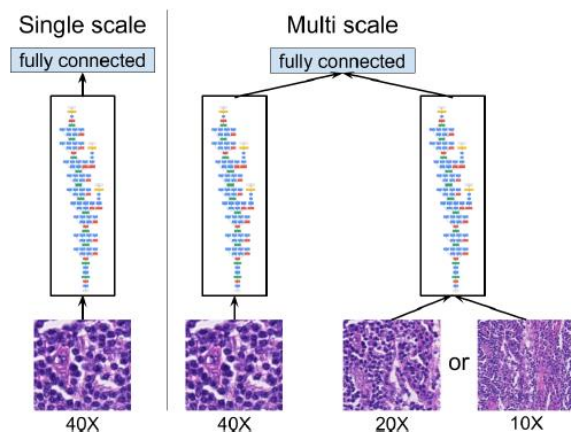
这周由于自身原因没能按时的去睿医出席，只是在玉泉校区进行学习以及一些知识上的积累。继续对于卷积神经网络的简化进行研究。对于之前确定的几个方法的试验也已经开始，减少卷积核的试验，对于我所选的神经网络没能成功，虽然提高了速度，但是性能减少的太多。对于全连接层使用循环结构的试验，还在进行中。选用了传统的 alexnet 作为例子。

## To do

接下来的一周，会去做好出席和继续深化自己的试验和学习。

## Data-free Parameter Pruning for Deep Neural Networks

本文提出了减少训练深度神经网络参数的方法。并非是像前人所做的那样减少权重，而是移除神经元。文章展现了如何找出相似的神元，并如何系统的去除它。作为例子，期在MINST数据集上进行了测试，减少了85%的参数，影响了35%的性能。



**Fig. 3.** The three colorful blocks represent Inception (V3) towers up to the second-last layer (PreLogit). *Single scale* utilizes one tower with input images at 40X magnification; *multi-scale* utilizes multiple (*e.g.*, 2) input magnifications that are input to separate towers and merged.

## Quantized Convolutional Neural Networks for Mobile Devices

高昂的硬件成本和巨大的计算空间限制了CNN的进一步扩展。本文，作者提出一种名为Quantized CNN的框架来加速计算以及减少CNN的存储和记忆空间。文章既使用了内核过滤也对全连接层的权重矩阵进行了量化。文章最后还在手机上进行了试验。

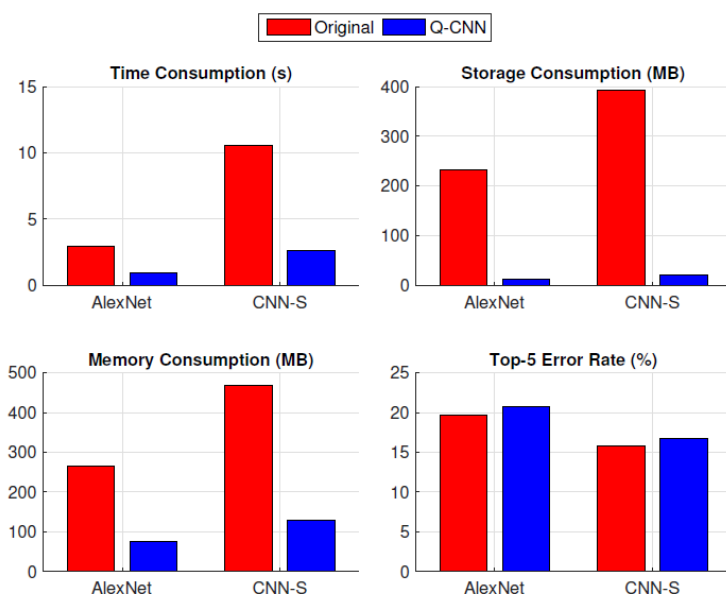


Figure 1. Comparison on the efficiency and classification accuracy between the original and quantized AlexNet [16] and CNN-S [1] on a Huawei® Mate 7 smartphone.

## A Kernel Redundancy Removing Policy for Convolutional Neural

### Network

文章给出了一种减少CNN种卷积核冗余的思路。CNN经常采用过参数化的卷积核来提取特征。本文则主张在不同的层使用不同的阈值来减少运算量。本文给出了一种关于冗余的定义并给出减少的策略，并对此进行了初步的试验。根据一定的阈值来过滤稀疏度较高的卷积核，进而精简CNN网络结构，提高模型运行效率。作者将文献《Accurate image super-resolution using very deep convolutional networks》中使用的残差卷积网络作为测试网络，并以卷积网络输出在峰值信噪比(Peak Signal to Noise Ratio, PSNR)上的损耗作为模型性能的评估标准。试验中在只减少1%的表现的情况下实现了50%的运算量缩减。